基于 VGGNet 卷积神经网络的加密芯片模板攻击新方法 *

郭东昕 [1]，陈开颜 [1†]，张 阳 [1]，张晓宇 [2]，李健龙 [1]

(1. 陆军工程大学石家庄校区 装备模拟训练中心，石家庄 050003; 2. 解放军 78090 部队，成都 610036)

摘 要：针对传统模板分析在实际攻击中的难解问题，重点研究了在图像识别领域具有优异特征提取能力的 VGGNet 网络模型，提出了一种基于 VGGNet 网络模型的模板攻击新方法。为了防止信号质量对模型准确率带来较大影响，采用相关性能量分析方法对采集到的旁路信号质量进行了检验；为了适应旁路信号数据维度特征，对网络模型结构进行适度调整；在网络训练的过程中，对梯度下降速率较慢、梯度消失、过拟合等问题进行了重点解决，并采用五折交叉验证的方法对训练好的模型进行验证。最终实验结果表明，基于 VGGNet 模型的测试成功率为 92.3%，较传统的模板攻击效果提升了 7.7%。

关键词：VGGNet 模型；加密芯片；模板攻击；能量分析

中图分类号：TP393.08　　doi: 10.3969/j.issn.1001-3695.2018.04.0255

New template attack method for encryption chip based on VGGNet convolutional neural network

Guo Dongxin[1], Chen Kaiyan[1†], Zhang Yang[1], Zhang Xiaoyu[2], Li Jianlong[1]

(1. *Shijiazhuang Campus of the Army Engineering University*, *Shijiazhuang* 050003, *China*; 2. *Unit* 78090 *of PLA*, *Chengdu* 610036, *China*)

Abstract: In order to solve the difficult problem of traditional template analysis in practical attacks, this paper focuses on the VGGNet network model with excellent feature extraction capabilities in the field of image recognition. We propose a new template attack method based on VGGNet network model. In order to prevent the signal quality from affecting the accuracy of the model, we used the correlation power analysis method to test the quality of the collected side-channel signal; In order to adapt to the dimensional characteristics of the side-channel signal data, We make modest adjustments to the network model structure; In the process of network training, we have focused on issues such as slower gradient descent, gradient disappearance, and overfitting, and we use a five-fold cross validation method to validate the trained model. The final experimental results show that the test success rate based on the VGGNet model is 92.3%, which is 7.7% higher than the traditional template attack effect.

Key words: VGGNet model；encryption chip；template attack；power analysis

0 引言

当前传统的模板攻击[1]在实际攻击中存在特征点数量选取限制以及矩阵求逆的数值问题，基于神经网络的模板分析在特征点数量选择方面不受限制，并且可以有效避免矩阵求逆问题。

同时近年来，深度学习领域发展较为迅猛，提出了较多优秀的特征提取网络模型，本文是为了提高模板攻击效率，采用 2014 年提出的 VGGNet[2]深度神经网络模型，相较于传统的基于高斯分布的模板攻击，由于 VGGNet 模型具有深的网络层次，更好地特征提取能力。在对其结构进行适应性调整后，将其应用到旁路模板分析领域，从而提高模板攻击的效率。

本文主要工作包括：a)在阐述了卷积神经网络模型的实现原理的基础上，详细分析了 VGGNet-16 模型的网络结构、参数配置，并对网络进行了适应性的结构调整;b)为了确保使用的旁路信号质量对构建的模型产生较大的影响，对信号质量进行了验证;c）在 VGGNet-16 网络模型训练过程中，针对网络模型的欠拟合问题，对训练数据进行了多项式扩充变换，增加数据特征，扩大不同类别数据特征差异；针对过拟合问题引入了 dropout 技术以及 L2 正则化项；针对梯度下降速度较慢，训练周期较长这一问题，对原始数据进行归一化处理;d）在使用同一设备同一批次采集数据的情况下，基于高斯分布的模板分析[3]与基于 VGGNet 模型的模板分析进行了对比实验，实验结果

收稿日期：2018-04-18；**修回日期：**2018-06-06　　**基金项目：**国家自然科学基金资助项目（51377170）；国家青年科学基金资助项目（61602505）

作者简介：郭东昕（1993-），男，山西山阴人，硕士研究生.主要研究方向为密码安全、旁路攻击；陈开颜（1970-），女（通信作者），辽宁盖县人，副教授，硕导，主要研究方向为密码学；张阳（1984-），男，河北南宫人，主要研究方向为系统安全；张晓宇（1992-），男，甘肃金昌人，硕士，主要研究方向为旁路攻击；李健龙，（1993-），男，吉林长春人，硕士研究生，主要研究方向为大数据分析.

表明新方法具有更高地匹配成功率，提升了 7.7%。

1　概念描述

1.1　类别特征转换

类别特征转换即将非数值型特征类别转换为数值型的特征类别。本文中涉及到将 9 种不同的汉明重量模型转换为数值型的类别表示，以便于统一数据标签格式和方便机器训练。

由于共有九种不同类型的数据，分别按照汉明重量 0~8 的顺序进行编号，编号为 1~9。将每个类别使用维度为 9 的向量表示，向量中类别对应的编号位置 1，其余为置 0，具体表示如表 1 所示。

表 1　数据类别转换

数据类别	数据标签
HW=0	100000000
HW=1	010000000
HW=2	001000000
HW=3	000100000
HW=4	000010000
HW=5	000001000
HW=6	000000100
HW=7	000000010
HW=8	000000001

1.2　多项式特征

多项式特征主要是通过将特征数据按照规定好的多项式法则进行运算，从而得到更多的特征属性，有利于更好地进行类别划分。

假设输入的特征属性为（a1,a2）,多项式最高次为 2，则得到的多项式特征数据为（1，a1,a2,a1^2,a1^a2,a2^2）。

在网络模型训练过程中，通过使用多项式特征数据，增加了不同类别数据的差异度，较好地解决了模型训练中遇到的欠拟合问题，具体示例如图 1 所示。

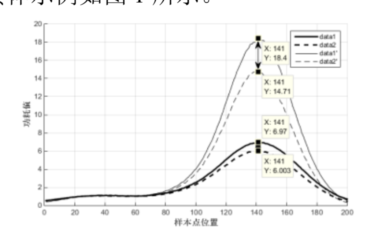


图 1　多项式特征

图 1 中 data1 与 data2 为任选的相邻汉明重量的功耗轨迹，data1’与 data2’是对原始数据进行多项式特征变化得到的数据。在点 X=141 处，data1 与 data2 的功耗值有最大差值，为 0.97，而经过变换后的最大差值为 3.69。显著增加了不同类别的特征区分度，更有利于模板快速高效地构建。

1.3　随机梯度下降

首先介绍一下与随机梯度算法[4]有关的定义与概念。

定义 1　假设函数 f 在点 a(x,y,z)处沿方向 l(方向角为 α 、β 、γ)存在下列极限:

$$\lim_{\rho\to 0}\frac{\Delta f}{\rho}=\lim_{\rho\to 0}\frac{f(x+\Delta x,y+\Delta y,z+\Delta z)-f(x,y,z)}{\rho} \quad (1)$$

记做 $\frac{\partial f}{\partial l}$，则称 $\frac{\partial f}{\partial l}$ 为函数在点 p 处沿方向 1 的方向导数。

实际上梯度在方向 1 处的投影即为函数在点 p 处沿着 1 方向的方向导数。方向导数公式为

$$\frac{\partial f}{\partial l}=\frac{\partial f}{\partial x}\cos\alpha+\frac{\partial f}{\partial y}\cos\beta+\frac{\partial f}{\partial z}\cos\gamma \quad (2)$$

令向量：

$$\vec{G}=\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y},\frac{\partial f}{\partial z}\right) \quad (3)$$

$$\vec{l}=(\cos\partial,\cos\beta,\cos\gamma) \quad (4)$$

有

$$\frac{\partial f}{\partial l}=\vec{G}\cdot\vec{l}=\left|\vec{G}\right|\cos(\vec{G},\vec{l})(\left|\vec{l}\right|=1) \quad (5)$$

当 $\vec{l}$ 与 $\vec{G}$ 方向相同时，余弦值为 1，方向导数取最大值：

$$\max(\frac{\partial f}{\partial l})=\left|\vec{G}\right| \quad (6)$$

$\vec{G}$ 向量的方向表示函数 f 变化率最大的方向，向量的值表示函数 f 的最大变化率之值。

定义 2　设函数 z=f(x,y)在平面区域 D 内具有一阶连续偏导数，则对于每一个点 P（x,y）∈D,向量：$\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right)$ 为函数 z=f(x,y)在点 P 的梯度，记做 gradf(x,y)，即

$$gradf(x,y)=\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right) \quad (7)$$

将式(7)中的向量转换为计算式得式（8）。

$$\mathrm{grad}f(x,y)=\frac{\partial f}{\partial x}\vec{i}+\frac{\partial f}{\partial y}\vec{j} \quad (8)$$

梯度为矢量单位，表示在函数中该点的方向导数在该方向上为最大值。即从该点处出发，沿着梯度的方向变化最快，快化率最大。

$\nabla=\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y}\right)$ 称为奈布拉（Nebla）算符，则梯度可简化为

$$gradf=\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right)=\nabla f \quad (9)$$

梯度下降法也称最速下降法，是用来在模型拟合数据的过程中通过不断地迭代从而得到最优值的优化算法。下降是指迭代方向为梯度的相反方向，即沿着梯度的相反方向搜索，离目标值越近，步长越小，速度越慢。

$$x_{k+1} = x_k + \alpha_k s_k \quad (10)$$

式（10）为梯度下降法的迭代算法，其中 x 为需要求解的值，s 为梯度负方向，α 为步长。

随机梯度下降算法事实上是梯度下降算法的一个扩展，在深度学习中的代价函数可以分解成每个样本的代价函数的总和。因此随着训练集规模增长，计算一步梯度也会消耗很长的时间。

随机梯度下降的核心是，梯度是期望。期望可使用小规模的样本近似估计。具体而言，在算法的每一步，我们从训练集中均匀抽出一小批量样本，小批量数据数目相对于总训练量较小。梯度的估计可以表示成

$$g = \frac{1}{m}\nabla_\theta \sum_{i=1}^{m} L\left(x^{(i)}, y^{(i)}, \theta\right) \quad (11)$$

其中：m 为小批量样本数，L 为每个样本的损失，使用来自小批量的样本，然后，随机梯度下降算法使用如下的梯度下降估计：

$$\theta \leftarrow \theta - \varepsilon g \quad (12)$$

其中：ε 是学习率，g 为梯度的估值。

1.4 欠拟合与过拟合

若给定的模型在特定的训练集中无法获得较低的误差值，则属于模型欠拟合训练数据集。若训练误差与测试误差存在较大差距，则属于模型过拟合训练数据集。

为了定量评估模型为过拟合还是欠拟合，引入了模型容量的概念，即指特定模型拟合各种函数的能力。一般来说，容量越低的模型拟合训练集的难度越大，容量越高的模型拟合训练集的难度越小。

在网络模型训练过程中通过增加特征点，对特征点进行多项式预处理等手段来解决模型欠拟合问题，并使用正则化来解决模型的过拟合问题。

1.5 正则化与 dropout 技术

正则化提出的目的是用来解决模拟过拟合问题，即降低模型的复杂度。具体是在损失函数中添加一个正则项的方式来降低过拟合。正则项主要有 L1 正则项和 L2 正则项两种。

$$L1 = \alpha \|w\|_1 \quad (13)$$

$$L2 = \alpha \|w\|_2 \quad (14)$$

其中：α 表示正则化系数，w 表示模型中的权值。L1 正则化表示模型中的权值绝对值之和，L2 正则化表示模型权值平方根下的平方和。

Dropout 技术也属于正则化，主要是通过人为设定比例随机去掉神经网络隐藏层中的神经元来解决模型过拟合问题，每次只使用部分神经元来训练模型中的权重和偏置。

通常情况下，基于对于同一个训练数据集，不同的神经网络训练之后，求得其输出的平均值，通常可以减少过拟合现象的发生。Dropout 技术就是利用这一原理，每次迭代训练时，随机去掉一部分神经元，即每次使用结构不同的模型对数据进行拟合。有效减弱了每层神经元之间的相互依赖关系，从而使训练后的神经网络模型具有更强的泛化能力。

因此除用于解决过拟合问题之外，较强的泛化能力可以增强分类的准确度，具体的实现原理如图 2 所示。

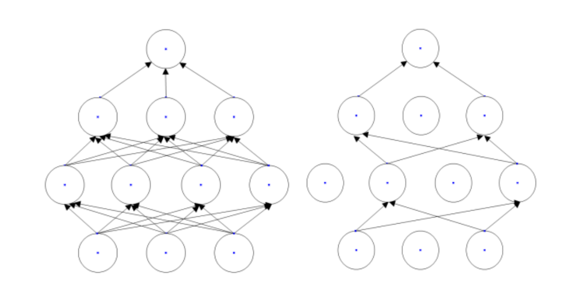


图 2 dropout 技术的实现原理

图 2 左侧为完整的神经网络结构，右图为加入 dropout 技术后，网络在迭代过程中按设定的比例丢掉隐藏层中的神经元。

2 模型分析与调整

2.1 VGGNet 神经网络模型介绍

Karen Simonyan 和 Andrew Zisser -man 提出的 VGGNet 网络模型[2]在 2014 年的 ILSVRC 竞赛中取得了优异的成绩，拥有较好地特征提取能力，该网络是从 Alexnet 网络发展而来的，网络结构主要的变化是采用了更小的卷积核以及更多的网络层数。竞赛结果表明，VGGNet 网络模型具有更优异的特征获取能力。具体模型如图 3 所示。

VGGNet网络模型结构					
A 11层	A-LRN 11层	B 13层	C 16层	D 16层	E 19层
输入层					
3*3-64	3*3-64 LRN	3*3-64 3*3-64	3*3-64 3*3-64	3*3-64 3*3-64	3*3-64 3*3-64
最大池化层					
3*3-128	3*3-128	3*3-128 3*3-128	3*3-128 3*3-128	3*3-128 3*3-128	3*3-128 3*3-128
最大池化层					
3*3-256 3*3-256	3*3-256 3*3-256	3*3-256 3*3-256	3*3-256 3*3-256 1*1-256	3*3-256 3*3-256 3*3-256	3*3-256 3*3-256 3*3-256 3*3-256
最大池化层					
3*3-512 3*3-512	3*3-512 3*3-512	3*3-512 3*3-512	3*3-512 3*3-512 1*1-512	3*3-512 3*3-512 3*3-512	3*3-512 3*3-512 3*3-512 3*3-512
最大池化层					
3*3-512 3*3-512	3*3-512 3*3-512	3*3-512 3*3-512	3*3-512 3*3-512 1*1-512	3*3-512 3*3-512 3*3-512	3*3-512 3*3-512 3*3-512 3*3-512
最大池化层					
全连接层-4096					
全连接层-4096					
全连接层-1000					

图 3　VGGNet 网络模型结构

Karen Simonyan 和 Andrew Zisser -man 共设计了 A、A-LRN、B、C、D、E 六种网络模型，对比实验表明，模型 D 效果最好，本文采用的 VGGNet 网络模型就是 D 模型，共有 16 层。D 模型主要分为 13 层卷积层和 3 层全连接层，为了强化特征提取效果，每个卷积核都选用 3*3 大小，由图 1 可知，首先为 13 个卷积层：第一、第二层卷积层各有 64 个特征图，第三、第四层卷积层各有 128 个特征图，第五、第六、第七层卷积层各有 256 个特征图，第八、第九、第十层、第十一、十二、十三层卷积层各有 512 个特征图。五个子采样层分别位于第二层卷积与第三层卷积、第四层卷积与第五层卷积、第七层卷积与第八层卷积、第十层卷积与第十一层卷积、第十三层卷积与第一层全连接层之间，缩放因子都为 2。最后是三个全连接层，第一层与第二层分别有 4096 个神经元，第三层的神经元个数由分类数目决定，这里是 1000 个图像类别，因此设置为 1000 个神经元。

2.2　VGGNet 模型结构调整

为了将 VGGNet 模型应用于旁路数据分析，对模型做以下调整，首先由于旁路信号为一维数据不同于图片的二维数据，也没有数据通道，因此需要将所有的卷积核的视野大小由 3*3 设置为 2*1，将卷积核缩小是为了进一步提高模型特征敏感度，下采样层在处理数据过程中也需要将二维处理区域转为一维。其次，旁路信号特征差异较小，为了加快模型建立，提高模型对旁路数据分类的准确率，将旁路数据按汉明重量模型[5]分为 9 类，因此需要将模型输出层的神经元个数设置为 9。具体调整后的网络结构如图 4 所示。

调整后的VGGNet网络模型结构
D
16层
输入层
2*1-64 2*1-64
最大池化层
2*1-128 2*1-128
最大池化层
2*1-256 2*1-256 2*1-256
最大池化层
2*1-512 2*1-512 2*1-512
最大池化层
2*1-512 2*1-512 2*1-512
最大池化层
全连接层-4096
全连接层-4096
全连接层-9

图 4　VGGNet 网络模型结构调整

3　实验

本实验对 AT89C52 微控制器写入 AES-128 加密算法，在 AES 第一轮的轮密钥加操作加入触发，加密过程中通过 RS-232 串口向其传输随机明文，通过 TeKtronix DPO4032 示波器采集旁路功耗信号。

3.1　旁路信号质量评估

在网络模型训练过程中，为了确保不出现因数据信号质量[6]问题而导致模型欠拟合，数据的验证和清洗是必不可少的，在整个实验过程中占有重要的地位。本文针对 AES 算法的前 8 位密钥进行破解，使用相关性能量分析方法分别对 15 条、50 条、100 条、200 条旁路功耗数据进行了攻击，如图 5~8 所示。

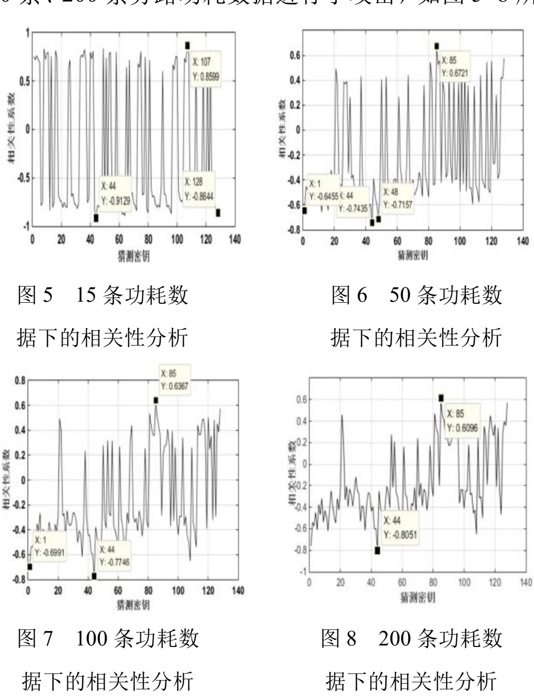


图 5　15 条功耗数据下的相关性分析　　图 6　50 条功耗数据下的相关性分析

图 7　100 条功耗数据下的相关性分析　　图 8　200 条功耗数据下的相关性分析

由图所示，横坐标表示猜测密钥，共有 256 种可能，由于采用的汉明重量模型在 8 位密钥中存在对称性[7]，因此只需要展示 128 种可能即可。纵坐标表示猜测密钥对应的相关性系数。随着功耗轨迹数量的增加，非正确密钥的相关性系数明显下降，正确密钥设置为 0x2B，也就是图中标出的 X：44，在所有猜测密钥中相关性系数最大，且具有较好地区分度。因此，通过实验验证，模型所使用的功耗信号质量较好。

3.2　汉明重量模型验证

为了确保汉明重量模型在 AT89 -C52 中的数据相关性，在相关性能量分析中正确密钥对应的相关性最大点处进行轨迹划分，攻击点为第一轮轮密钥加，将前 8 位明文与对应密钥进行轮密钥加操作，对不同操作结果中间值对应的汉明重量进行划分，共分为 9 中，分别为 0~8，如图 9 所示。

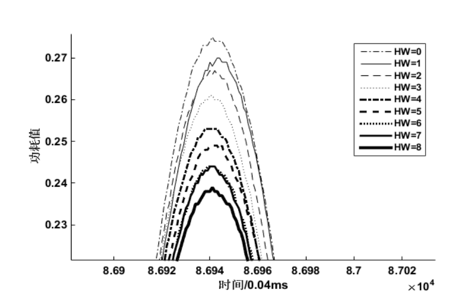


图 9　不同汉明重量对应的功耗轨迹

在采集数据过程中，每条能量轨迹采集了 100000 个点，点数较多，需要在构建旁路模板之前，对数据能量相关性最大位置进行定位，本文使用相关性能量分析方法（CPA）进行定位，如图 10 所示。

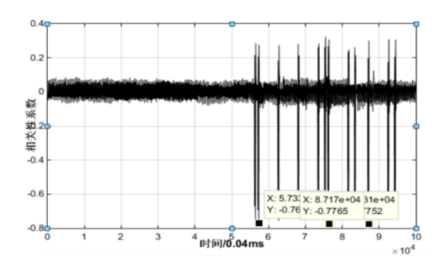


图 10　相关性最大点位置

在多个位置均出现较大相关性，在特征点提取时，以相关系数的值作为指标，按照由大到小的次序对值所对应点的时刻进行记录，从而确保提取到最优特征点。

3.3　对比实验

本文分别使用基于高斯分布的模板攻击方法[9]、基于 VGGNet 网络模型的模板攻击[2]进行对比。为了确保对各种方法的准确评估，实验使用的数据为同一设备同一批次采集。训练数据共 9 000 条，每个汉明重量 1 000 条，测试数据共 1 800 条，每个汉明重量 200 条。为保证测试结果准确，测试数据在模型训练过程中不会使用。在神经网络[10][11]训练过程中使用五折交叉验证法进行成功率检验，即将训练数据等分为五份，各 1 800 条，在模型训练阶段，依次作为验证集，其余数据作为训练集，对模型进行验证，模型训练完成后，再使用测试集对训练好的模型进行测试。

实验过程中通过 3.2 节提到的方法选取 100 个特征点，为了增大不同汉明重量类别数据的特征，采用 2.2 节中的多项式特征方法对特征点进行平方预处理，较好地解决了模型欠拟合问题。具体的实验结果如图 11 与表 2 所示。

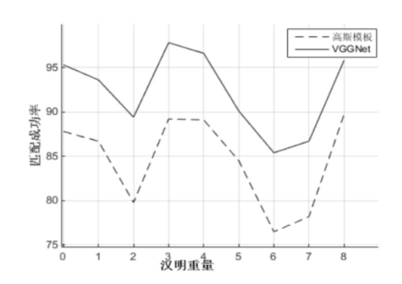


图 11　不同汉明重量类别匹配成功率

表 2　模板攻击实验结果

分类方法	评估指标		
	平均匹配成功率	攻击耗时/min	训练耗时/h
基于高斯分布的模板攻击	84.6%	1.65	0.6
基于 VGGNet 的模板攻击	92.3%	3.07	36

从图 11 中可知，整体上基于 VGGNet 网络模型的模板攻击的匹配成功率高于基于传统高斯分布的模板攻击。在汉明重量为 1 和 2 与 6 和 7 的匹配率较低，说明采集到的数据在这些汉明重量之间的区分度较低，在模型进行分类任务时，造成较低的匹配成功率。

从表 2 中可知，由于传统的模板攻击在模板构建过程中不需要进行模板参数调整，因此模板构建所需时间较少，而基于 VGGNet 模型的模板攻击在训练过程中需要不断地迭代更新模型参数，因此耗时较多，若在训练过程中出现过拟合问题则需要更多时间调整网络结构或添加正则化项重新训练。两种方法在攻击耗时方面相差不大。

在平均匹配成功率方面，基于高斯分布的传统模板攻击的平均匹配成功率为 84.6%，低于基于 VGGNet 模型的模板匹配成功率。新方法较传统的模板攻击成功率有较大提高。

4　结束语

本文提出了基于 VGGNet 网络模型的模板攻击新方法，将改进后的 VGGNet 深度神经网络应用于旁路分析领域，利用深度神经网络具有较强的特征提取能力这一特点，提高了旁路模板攻击的匹配成功率。通过采集在 AT89S52 单片机上运行的 AES-128 加密算法的功耗旁路信号，并对同一批数据进行了对比试验，结果表明：基于 VGGNet 模型的模板攻击匹配正确成功率高于传统的模板攻击，较传统的模板攻击提高了 7.7%。

参考文献：

[1] Kocher P C. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems [C]// Proc of CRYPTO. 1996: 104-113.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014.

[3] Kocher P C, Jaffe J, Jun B, *et al.* Introduction to differential power analysis [J]. Journal of Cryptographic Engineering, 2011, 1, (1): 5-27.

[4] Goodfellow I, Bengio Y, Courville A. 深度学习 [M]. 赵申剑，黎彧君，符天凡，等译. 北京：人民邮电出版社, 2017. (Goodfellow I, Bengio Y, Courville, Deep learning [M]. Zhao Shenjian, Li Yijun, Fu Tianfan, *et al.* Beijing: People's Posts and Telecommunications Press, 2017.)

[5] 欧长海，王竹，黄伟庆，等. 基于汉明重量模型的密码设备放大模板攻击 [J]. 密码学报, 2015, 2 (5): 477-486. (Ou Changhai, Wang Zhu, Huang Weiqing, *et al.* Hamming weight model based cryptographic device amplification template attack [J]. Journal of Ccitonium, 2015, 2 (5): 477-486.)

[6] 张晓宇，陈开颜，张阳，等. 基于 DTW 算法的旁路功耗信号动态伸缩

对齐 [J]. 计算机应用研究, 2017, 34 (9): 2782-2785. (Zhang Xiaoyu, Chen Kaiyan, Zhang Yang, *et al.* Dynamic telescopic alignment of bypass power consumption signal based on DTW algorithm [J]. Application Research of Computers, 2017, 34 (9): 2782-2785.)

[7] 张阳, 陈开颜, 李雄伟, 等. 基于差异度的密码芯片旁路攻击研究 [J]. 通信学报, 2015, 36 (3): 2015066. (Zhang Yang, Chen Kaiyan, Li Xiongwei, *et al.* Research on password chip bypass attack based on difference degree [J]. Journal of Communications, 2015, 36 (3) .)

[8] Mangard S, Oswald E, Popp T. 能量分析攻击 [M]. 冯登国 等译. 北京: 科学出版社, 2010. (Mangard S, Oswald E, Popp T. Energy analysis attacks [M]. Feng Dengguo *et al.* Beijing: Science Press, 2010.)

[9] 邓高明, 赵强, 张鹏, 等. 针对密码芯片的电磁频域模板分析攻击 [J]. 计算机学报, 2009, 32 (4): 602-610. (Deng Gaoming, Zhao Qiang, Zhang Peng, *et al.* Electromagnetic domain template analysis attacks on cipher chips [J]. Chinese Journal of Computers, 2009, 32 (4): 602-610.)

[10] 余凯, 贾磊, 陈雨强, 等. 深度学习的昨天、今天和明天 [J]. 计算机研究与发展, 2013, 50 (9): 1799-1804. (Yu Kai, Jia Lei, Chen Yuqiang, *et al.* Deep learning, yesterday, today and tomorrow [J]. Journal of Computer Research and Development, 2013, 50 (9): 1799-1804.)

[11] Cai Z, Fan Q, Feris R S, *et al.* A unified multi-scale deep convolutional neural network for fast object detection [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 354-370.